BMC Medical Research Methodology

BioMed Central

Software

Open Access

The rating reliability calculator

David J Solomon*

Address: Office of Medical Education Research and Development and the Department of Medicine, Michigan State University, East Lansing, Michigan, USA

Email: David J Solomon* - dsolomon@msu.edu

* Corresponding author

Published: 29 April 2004

BMC Medical Research Methodology 2004, **4**:11

Received: 31 January 2004

Accepted: 29 April 2004

This article is available from: http://www.biomedcentral.com/1471-2288/4/11

Abstract

Background: Rating scales form an important means of gathering evaluation data. Since important decisions are often based on these evaluations, determining the reliability of rating data can be critical. Most commonly used methods of estimating reliability require a complete set of ratings i.e. every subject being rated must be rated by each judge. Over fifty years ago Ebel described an algorithm for estimating the reliability of ratings based on incomplete data. While his article has been widely cited over the years, software based on the algorithm is not readily available. This paper describes an easy-to-use Web-based utility for estimating the reliability of ratings based on incomplete data using Ebel's algorithm.

Methods: The program is available public use on our server and the source code is freely available under GNU General Public License. The utility is written in PHP, a common open source imbedded scripting language. The rating data can be entered in a convenient format on the user's personal computer that the program will upload to the server for calculating the reliability and other statistics describing the ratings.

Results: When the program is run it displays the reliability, number of subject rated, harmonic mean number of judges rating each subject, the mean and standard deviation of the averaged ratings per subject. The program also displays the mean, standard deviation and number of ratings for each subject rated. Additionally the program will estimate the reliability of an average of a number of ratings for each subject via the Spearman-Brown prophecy formula.

Conclusion: This simple web-based program provides a convenient means of estimating the reliability of rating data without the need to conduct special studies in order to provide complete rating data. I would welcome other researchers revising and enhancing the program.

Background

Rating scales form an important means of gathering evaluation data in medical education, health services research and quality assurance. Since ratings are often used in making high stakes decisions, assessing the reliability of ratings scales can be critically important. A variety of techniques have been developed for estimating the reliability of ratings. When ratings are in the form of numerical scales, reliability is most commonly assessed using techniques based on classical test theory [1] or more recently, an extension of it, generalizability theory [2].

Generalizability theory, which is based on the analysis of sources of variance in the ratings, provides a powerful tool not only for estimating the reliability of ratings but also for developing efficient designs for obtaining reliable rat-

ings. Unfortunately generalizability studies require what is commonly termed a balanced design. Generally this means each subject being rated must be rated by the same set of judges. Without a balanced design, the estimation of the variance components upon which generalizability studies are based can be extremely complex and often without a clear solution [3].

When ratings are collected in real-world situations, the result is rarely a balanced design. For example, when faculty preceptors rate students and residents, different trainees are generally rated by different faculty and often by different numbers of faculty. A similar situation occurs when patients provide feedback on their physicians or other health providers.

While it is often possible to design studies that capture complete rating data in order to conduct a generalizability study, this can be expensive and time consuming. Often it is only practical to collect small sets of complete ratings which are susceptible to sampling error.

Over 50 years ago Robert L. Ebel wrote a classic article discussing techniques for estimating the reliability of ratings based on incomplete data [4]. He described an algorithm which is also based on analysis of variance and approximates an intraclass correlation, but that requires no assumptions about the number of judges or whether the same judges rate each subject. Though Ebel's article predates the development of generalizablilty theory and his approach lacks its power and flexibility, his approach provides a simple means of deriving a reasonable estimate of the reliability of ratings based on incomplete data. The algorithm can be easily calculated using data collected in the normal course of evaluating students in educational settings, health care providers by their patients or health services by the consumers of those services without the need for designing and conducting specialized generalizability studies.

Ebel's article has been widely cited over the years but to my knowledge his algorithm for estimating the reliability of ratings is not available in any of the standard statistical software packages. The web-based software described below provides a convenient means to estimate the reliability of ratings based on Ebel's algorithm. It is available for public use on our server and the source code is available under GNU general public license [5].

Implementation

This simple program is written in a combination of HTML and PHP, a widely used open source imbedded scripting language. The software consists of a data entry form where the user specifies a data file with the rating data on their personal computer. The user can also specify the number of ratings that will be averaged to obtain a score. If more than one rating is specified, the program uses the Spearman-Brown prophesy formula [6] to extrapolate what the reliability would be if that number of ratings was averaged to obtain a score.

When the "submit" button on the data entry form is pressed, the data file is uploaded to the server and the resulting web page displays reliability estimate, number of subjects rated in the data set, the harmonic mean number of ratings per subject rated, mean of the ratings across all people in the data set and the standard deviation of the ratings. The program also lists each subject, their mean rating, standard deviation of their ratings and number of ratings they received.

The data file upload is accomplished by using the HTTP post method for upload which requires the user to have a RFC-1867 compliant browser. These include Netscape Navigator version 3 or later or a version of Windows Internet Explorer that is later the 3.0 [7].

The program checks for non-numeric ratings and subjects with only a single rating. Both of these issues can generate unpredictable results. The program provides a warning message and drops these ratings from the analysis.

Results

Rating data format

All that needs to be specified for each rating used to estimate the reliability is the numerical rating and an alphanumeric code that specifies the subject that is being rated. The data format is given below.

The ratings should be stored in an ASCII text file with each rating on a separate line. The rating should be preceded on the line by an alphanumeric identifier for the subject being rated. A comma "," should separate the alphanumeric identifier from the numeric rating. An example of how the file should look is shown below.

Joe, 5

Tom, 6

Joe, 6

Joe, 3

Sally, 7

Tom, 4

Sally, 8

In the example above, there are three ratings for Joe and two each for Tom and Sally. For the purposes of estimating the reliability via this approach, it is immaterial whether the same or different judges rated each of the three individuals. There is also no need to group the ratings for a particular subject together.

Creating the data file

You can create this file in a number of ways. If you are using a Windows-based computer, can to enter the data directly using Notepad. You could also use Word or another word processing program and save the file as a "text" file. It is also possible to create the data file using Excel. Use one column for the identifier and one for the rating, Once the file is created, save it as a comma delimited CSV file by specifying "CSV (comma delimited)" in the drop-down menu below where you enter the file name when you save the Excel workbook.

Uses

The Rating Reliability Calculator is appropriate for use where multiple judges rate each subject being rated using a scale that constitutes interval level measurement. Interval level measures constitute scales that increase monotonically where the intervals between adjacent scale values are equal with respect the attribute being measured [8]. There is no need for the same judges nor the same numbers of judges to rate each subject.

The algorithm treats variations in the stringency among the judges, e.g., the extent they are "hawks" versus "doves" as a source of error. In this sense, it produces what is sometimes termed "domain referenced" as apposed to"norm referenced" reliability coefficients [2]. Since there is no assumption made that the same judges rate each subject, domain referenced reliability coefficients are probably more appropriate.

Cautions

The reliability of ratings in theory ranges between zero and one. The algorithm used to estimate the reliability in this program can potentially generate in estimates of the reliability that are negative. This is true of any method of calculating reliability based on algebraically manipulating mean squares in order to obtain unbiased estimates of the sources of variance in the ratings. This generally means the actual reliability is near zero and the negative reliability estimate generated by the program is due to sampling error. If the reliability generated by the program is large and negative or you have good reason to believe the reliability of the ratings should be fairly high, check your data.

Conclusions

Although Ebel's approach to estimating rating reliability is over 50 years old, it continues to provide an extremely flexible method for estimating the reliability of ratings that continues to useful today.

Availability and requirements

Project name: The rating reliability calculator

Public use access: http://www.med-ed-online.org/rating/reliability.html

Source Code: http://www.med-ed-online.org/rsoftware.htm#Rating

Operating system: Platform independent

Other requirements: In order to use the public access version, the user must have a RFC-1867 compliant browser which includes Netscape Navigator version 3 or later or Windows Internet Explorer version later the 3.0. If the source code is installed on another server it must have PHP version 3 or higher installed and operating.

License: The software is available for use under the GNU General Public License http://www.gnu.org/copyleft/gpl.html. There are no other restrictions.

Competing interests

The author has no competing interests.

Authors' contributions

David J Solomon programmed the rating reliability calculator and wrote this article in its entirety. The conceptual frame work and algorithm was developed by Robert L. Ebel.

References

1. Magnusson D: *Test Theory* Reading, Mass: Addison-Wesley Publishing Co; 1966.
2. Bennan RL: *Generalizability Theory* New York: Springer-Verlag; 2001.
3. Bennan RL: *Generalizability Theory* New York: Springer-Verlag; 2001:228-231.
4. Ebel RL: **Estimation of the reliability of ratings.** *Psychometrika* 1951, **16:**407-424.
5. **Spearman-Brown prediction formula From Wikipedia, the free encyclopedia** [http://en.wikipedia.org/wiki/Spearman-Brown_prediction_formula]
6. **GNU General Public License** [http://www.gnu.org/copyleft/gpl.html]. accessed 01/30/04
7. **PHP Manual Chapter 18. "Handling file uploads"** [http://us2.php.net/manual/en/features.file-upload.php#features.file-upload.post-method]
8. **Scales of Measurement** [http://web.uccs.edu/lbecker/SPSS/scalemeas.htm]

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/4/11/prepub